Predicting the End-Price of Online Auctions

Rayid Ghani, Hillery Simmons

Accenture Technology Labs, Chicago, IL 60601, USA
Rayid.Ghani@accenture.com, Hillery.D.Simmons@accenture.com

Abstract. Online auctions have become one of the fastest growing modes of online-commerce transactions. eBay has 94 million active members buying and selling goods at a staggering rate. These auctions are also producing large amounts of data that can be utilized to provide services to the buyers and sellers, market research, and product development. We collect historical auction data from eBay and use machine learning algorithms to predict end-prices of auction items. We describe the features used, and several formulations of the price prediction problem. Using the PDA category from eBay, we show that our algorithms are extremely accurate and can result in a useful set of services for buyers and sellers in online marketplaces.

1 Introduction

Online marketplaces have been gaining in importance for the past few years. EBay, Yahoo! Auctions, Amazon Marketplace and other online marketplaces have become significant commercial entities and it is estimated that they will account for 25% of online ecommerce by 2005. Even today, eBay, one of the largest online marketplaces, consists of 94 million members typically offering 19 million items for sale at any given time. In 2003, $24 billion of goods were sold on eBay. Although such online marketplaces offer individuals a unique opportunity to buy and sell goods, they also offer a new source of data that can be utilized to make inferences about mass economic behavior, product marketing, market research, and provide services to the buyers and sellers participating in the online marketplaces.

In this paper, we describe our work on a system capable of predicting the end-price of auction listings. Price prediction for auctions is a challenging task for machine learning algorithms mainly because of the large number of attributes that can vary in auction settings. Even items (of the same kind) differ in condition. The variance in shipping charges, reliability of sellers, appearance of the listing, beginning and ending times, all are factors that make it difficult to predict the price of an auction. Even if all the above variations were accounted for, there is still the uncertainly in human behavior when bidding in auctions. Auction Software Review [1] reported that 15% of the auctions eBay

are won in the last minute which increases the uncertainty in the end-price of a given auction.

The price prediction system described in this paper is trained by using the characteristics of the seller, the item to be auctioned, the features of the auction, and historical auction data to make predictions about the outcome of an auction before it starts. We describe the features used, the various ways in which price prediction can be formulated as a machine learning problem, and the performance results of several algorithms applied to this task. These results show that we can predict the end-prices of auctions very accurately which leads to several applications that can be used to provide new services to the participants in online marketplaces.

2 Related Work

There has been a lot of work in the Economics and Data Mining community on analyzing online auctions. Most of this work has focused on describing past auctions rather than prediction. Bajari & Hortascu [2] develop econometric techniques to build models of bidding behavior. Lucking-Riely et al. [4] use data collected from eBay about auctions of collectible coins to study the factors that affect the price. Although this study is a good exploratory analysis of online auctions, it only finds correlations between attributes of the auction and the resulting price and does not aim to build predictive models.

There has been some work in price prediction of items in online markets e.g. airlines fares [5] but not much has been done in the auction domain. The only work we are aware of that involves predicting prices in auctions was done implicitly during the Trading Agent Competition (TAC) [9,11,12] focusing on the travel domain. TAC relies on a simulator of airline, hotel, and ticket prices and the competitors build agents to bid on these. TAC simulates prices and assumes that the supply of products (airline tickets) is unlimited. Several TAC competitors have explored a range of methods for price prediction including historical averaging, neural nets, and boosting. All of the work in this domain is performed with artificially generated data and does not use any real auction data. The work in this paper is based on data collected from eBay and is aimed at predicting the prices to provide a new set of services to the buyers and sellers in online marketplaces.

3 Overview of our Approach

At a high level, the initial goal of our work is to predict the ending price of a given auction before the auction starts. For the results presented later in this paper, we specifically deal

with eBay auctions but the algorithms and features should generalize to other online auctions. The input to the system is the data that is filled in by the seller when listing an item for auction. This includes information about the seller, details of the item (name, specifications, description, photos, etc.), and attributes about the auction (length, starting bid, reserve price, shipping charges, etc.). This information is processed to extract attributes and create new attributes that are then used to predict the probable end-price for that auction. The high-level steps of our approach are outlined below:

1. Collect data about auction listings
2. Define the set of features to be extracted
3. Create meta-features that are derived from the initial set of features
4. Train a classifier/extractor to use the training data to now extract features from unseen data

4 Data Collection

We constructed a web crawler to visit eBay and extract auction listings for several categories over a period of two months. For a given category, the crawler constructed a search query to find all completed auctions and stored all the pages associated with that auction. This included the page where the auction was listed in the search results, the detailed page for the auction containing the description, photos of the item, the bid history page containing usernames of all bidders, amount and time of all bids, as well as the page listing the feedback for the seller. For further analysis in this paper, we selected the PDA category. Table 1 gives some statistics about PDA auctions that were executed on eBay during February 2004.

5 Feature Extraction & Construction

The data collected by the crawler was then used to extract four classes of features:

1. Seller Features (Table2)
2. Item Features (Table 3)
3. Auction Features (Table 2 and 3)
4. Temporal Features

The temporal features were based on other auctions in the recent history for the same item. For each auction listing in our data set we extracted the following features:

Let X_{ij} be the set of auctions that finished in the j hours preceding the time when auction i started. For each auction item i, we constructed the set $\mathbf{X}$ consisting of sets X_{ij} with j having values 1/3,1/2,1,2,4,6,12,24. From each set in $\mathbf{X}$, we created new features consisting of the Mean, Standard Deviation, Minimum, Maximum values for the starting bid, shipping price and end price. We also calculated a feature counting the number of

Table 1. Statistics collected for auctions on EBay in the PDA category for February 2004

Total Count of Auctions (Volume)	51962
Total Count of Items in Auctions	491727
Total Count of Regular Auctions	44445
Sell Through Rate of Regular Auctions	20.70 %
Total Count of Reserve Auctions	6578
Total Count of Fixed Price Auctions	3437
Total Count of Auctions Ending with Buy It Now	7634
Total Count of Dutch Auctions	4949
Total Count of Items in Dutch Auctions	444714
Average Quantity Available in Dutch Auctions	89.86
Average Bids for each Auction	15.79
Average Bids for each Item in all Auctions	1.67
Average Bids per Auction (excluding auctions ending with Buy It Now)	16.67
Average Price for all Auctions	107.04
Average shipping amount (for items where shipping is listed)	12.12
Average Price for Auctions Ending with Buy It Now	154.87
Average Price for Single-Item Fixed Price Auctions	138.27
Total Bid Amount for all Auctions	5,572,895.68
Total Sales for all Auctions	5,043,139.16
Total Number of Sellers	19957
Average Feedback for Sellers	2,510.25
Average Percent Positive Feedback for Sellers	95.50 %

similar items that were listed for auction in the j hours before auction i started and the number of auctions where the item did not sell.

More formally, for each auction in our data set, we calculate the cross-product A x B x C where:

A={Mean, Standard Deviation, Minimum, Maximum} (measures to calculate)
B={Starting Bid, Ending Price, Shipping Charges} (features to use)
C={1/3,1/2,1,2,4,6,12,24} (number of hours preceding the starting of auction i)

These four kinds of features result in a total of approximately 430 features for each instance in our data set.

6 Price Prediction as a Machine Learning Problem

Given the features that were described in the previous section, the task now is to predict the end-price of a new auction. There are several ways in which this problem can be tackled with machine learning algorithms. We defined the problem in three ways to compare the relative merits of each approach:

Table 2. Features that were directly extracted from the auction listing

Feature	Description
TITLE	Auction title
SELLERRATING	Seller rating, e.g., assigned by the online marketplace based on feedback received by other online marketplace users
SELLERHASMEPAGE	Indicates the seller has a introductory/bio webpage on the online marketplace website
SELLERISPOWERSELLER	Indicates a seller has a large number of successful sales
FIRSTBID	The minimum price for the auction
ACCEPTSPAYMENTSERVICE	Indicates the seller accepts payments through a secure third party payment service
ISDUTCH	Indicates the auction is set up as a Dutch auction
ISRESERVE	Indicates the seller set up a reserve price for the auction
ISRESERVEMET	Indicates that the closing price exceeded the reserve price set by the seller
QUANTITYAVAILABLE	Indicates the number of items available
STARTDATE	The beginning date and time of the auction
ENDDATE	The ending date and time of the auction
ISFIXEDPRICE	Indicates the seller set up a "Buy it now" price for immediate sale of the item
SELLERHASSHADES	Indicates that the seller has recently changed their email and billing information
CATEGORY	The identifying number for the primary item category chosen for the auction
ISGIFT	Indicates the seller has chosen to add a gift box icon to the listing to indicate the item would be a good gift
SUBTITLE	Subtitle text if specified by seller
CATEGORY2	The identifying number for a secondary item category for the auction
PREFERS3RDPARTYPAYMENT	Indicates the seller's preferred method of payment is through a third party payment service.
POSITIVEFEEDBACKPERCENT	The percent of positive feedback (of all the feedback) received by the seller
HASPICTURE	Indicates the seller included a picture with the listing
MEMBERSINCE	The date the seller created their online marketplace user account
HASEBAYSTORE	Indicates the seller has an online retail page on the online site

Table 3. Derived Features used in our experiments

NEW	Indicates the existence of the word "new" in the title
BROKEN	Indicates the existence of the word "broken" in the title
LIKENEW	Indicates the existence of the phrase "like new" in the title
SEALED	Indicates the existence of the word "sealed" in the title
MANUFACTURER	The item manufacturer, extracted from the title
SCREEN	The item screen features, extracted from the title
MODEL	The item model, extracted from the title
MEMORY	The item memory features, extracted from the title
FEATURES	Other item features, extracted from the title
STARTDAYOFWEEK	The day of the week (number) that the auction started
ENDDAYOFWEEK	The day of the week (number) that the auction ended
AUCTIONLENGTH	The number of days that the auction lasted
BUYERPAYS	Contains "true" if buyer pays for shipping, "false" if seller pays
FREESHIPPING	Contains "true" if shipping is free to the buyer
SEARCHDESCRIPTIONFORSHIPPING	Indicates that the shipping amount was not specified in its designated place (ShippingAmount field) and a search was done in the description text to get the price
SHIPPINGCHARGE	The ShippingAmount or the amount found in the description text search

1. **Regression:** We treat the price prediction task as a regression task and use the training data to learn regression coefficients. The output of the model, when applied to new data is a specific (continuous) price. For the results reported in the following section, we used linear regression, polynomial regression with degrees 2 and 3, and CART (Classification & Regression Trees) .

2. **Multi-Class classification:** We discretize the end-price (target variable) into $5[1] intervals and create discrete categories. Each instance now falls in one of these categories. The price prediction problem can then be treated as a multiclass classification problem in which case the output is a $5 range instead of the specific

---

[1] The interval would vary with different tasks. We picked $5 because the average price of PDAs in our dataset was $55 and we wanted to predict the price within a 10% window of the average price.

price (as in the case of regression). We use decision trees (C5.0) and neural networks to implement the multiclass classification in our experiments.

3. **Multiple Binary classification tasks**: We create multiple binary classifiers, with each classifier learning a binary classification task: whether the end-price of the auction will be more than $X or not. For the experiments in this paper, we varied X in $5 intervals. For example, one classifier learned the task whether price is more than $5, the next for $10, and so on, going up to the maximum price in the training set.
   This technique was motivated by the small amounts of training examples that are available for any item in online auctions. Although there are a large number of auctions going on, auctions for any single kind of item are limited. This creates the need to use the scarce training data in an efficient manner. The multiclass classification scheme (described above) is not very effective at this goal since the positive examples for each category ($5 interval) are limited to the ones in that category. In contrast, for the binary classification case with multiple classifiers, the positive examples for the classifier that is predicting whether the price is going to be greater than $45, consist of all the examples where the price is greater than 45 (and not just in the range $45-$50). Each classifier has access to the entire training data instead of subsets that the multiclass classifier uses, making it much more effective at using the available training data. Our hypothesis is that this scheme will perform better than the multiclass classification in our evaluation. We use decision trees (C5.0) and Neural Networks to construct each classifier in this scheme.

7 Experimental Results

For our experiments, we selected all the auctions that were selling Palm Zire 21 from the PDA category on eBay during a two-month period. This resulted in a data set consisting of 1700 instances. The price distribution for these instances is shown in Figure 1. For evaluation, we used 1300 for training the models and the rest of the 400 for testing.

The results in Tables 4 and 5 show that all of the methods we use are effective at predicting the end-price of auctions. Regression results are not as promising as the ones for classification, mainly because the task is harder since an exact price is being predicted as opposed to a price range. In the future, we plan to narrow the bins for the price range and experiment with using classification algorithms to achieve more fine-grained results. Between the two schemes we used for classification (multiclass classification, and multiple binary classifiers), we see dramatic improvement from the second approach. We are able to achieve 96% accuracy by creating classifiers that learn separate binary classification tasks of predicting whether the price is more than $x for different values of

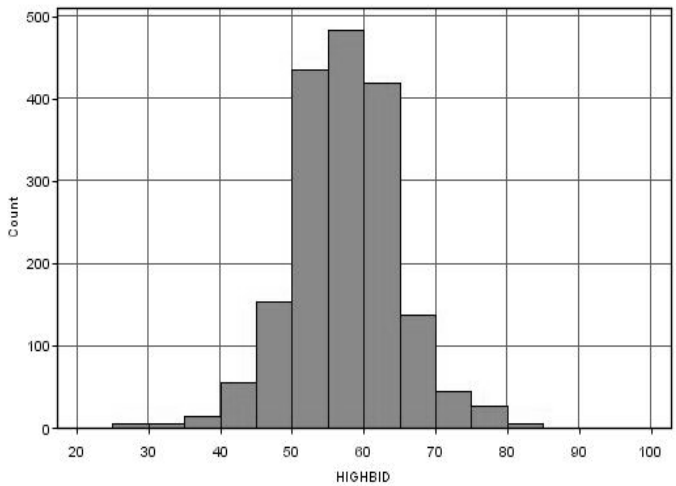

Fig. 1. End-Price Distribution for the data used in our experiments

Table 4. Results for Regression averaged over 5 random train-test sets with 1300 training examples, and 400 test examples. Baseline result is the mean of all the prices for examples in the test set.

	Mean Squared Error	Baseline
Linear Regression	5.9	6.6
CART (Regression Tree)	5.4	6.6

Table 5. Results for MultiClass and Binary Classifications averaged over 5 random train-test sets with 1300 training examples, and 400 test examples. Baseline is the category that would have been predicted by default (largest category)

	Accuracy	Baseline
MultiClass – C5.0	72%	26%
MultiClass – Neural Network	75%	26%
Binary Classifiers - C5.0	91%	26%
Binary Classifiers - Neural Networks	96%	26%

x. We believe that the improvement is consistent with our initial hypothesis that this technique utilizes all of the training data available with every classifier instead of being restricted to a particular category.

This idea has some similarity to the notion of using Output Codes for multiclass classification where a multiclass classification problem is decomposed into multiple binary problems with each classifier using all of the available training data [5,7].

8 Applications

The ability to predict the ending price of online auction items lends itself to a variety of applications. In this section, we briefly describe some concrete applications that we have developed.

Price Insurance: Knowing the end-price before an auction starts provides an opportunity for a third-party to offer price insurance to sellers. The insurer, knowing the likely ending price for any auction listing before it starts, can charge a premium to insure that the item will sell for at least the insured price. If the item sells for less than the insured price, the seller is reimbursed for the difference (between the insured price and the selling price) by the insurer. We have done some simulations using the price prediction algorithms described in this paper and have found that this insurance service would be profitable given the accuracy of the price prediction algorithms. We are currently in the process of doing detailed experiments and simulations with the price insurance algorithms.

Listing Optimizer: The model of the end-price based on the input attributes of the auction can also be used to help sellers optimize the selling price of their items. When the seller enters their personal information and the item they want to sell in an auction, our service would give suggestions for the auction attributes (such as starting time, starting bid, use of photos, reserve price, words to describe the item, etc.) that would maximize the end-price.

There are several other applications that can be enabled by the price prediction techniques described in this paper. While we have not provided an exhaustive list of applications, we believe that having access to the likely end-price of auction items opens up a large variety of services that can be offered to both buyers and sellers in online auctions.

9 Conclusions and Future Work

We described our work on a system capable of predicting the end-price of online auctions. The system requires the information provided by the seller of an item and uses machine learning algorithms to accurately predict the end-price. We find that among a variety of problem formulations, posing price prediction as a series of binary classification problems is best suited for this task. There are several ways to extend the applicability of our approach and try alternative methods. In this paper, we use PDAs because they can be described and compared using "hard" features/specifications (e.g memory size, speed, screen type, operating system). In contrast, "soft" products such as clothing items don't have the same kinds of attributes that can be used to compare different kinds of items. Features such as size, material and color do exist but they are not the kind of attributes that "define" the style of the product. To apply the algorithms in that context, we can use ideas described in some earlier work [8] to first extract product attributes from free-text descriptions of products available online (in stores or auction websites), and then use these attributes as part of the learning process. This would extend the applicability of our approach to "soft" products such as apparel, fashion items, antiques, and collectibles.

In this paper, we only used data from auctions that were about the same item. We encoded the context by using temporal features that described past auctions that were "similar" to the one that was being studied. Another direction that we intend to follow is to use data about auctions that are not related to the current item. This is similar to work done in machine learning from learning with unlabeled data [3,10] where the unlabeled data implicitly provides background knowledge and correlations between attributes that are not directly related, but useful for the classification task. Since there is data available for auctions in general which can be collected fairly cheaply, it would be valuable to study and develop techniques that can learn general patterns about auctions to make inferences about specific items and auctions.

References

1. Auction Software Review. http://www.auctionsoftwarereview.com/article-ebay-statistics.asp
2. Bajari, P. and A. Hortacsu, "Winner's Curse, Reserve Prices, and Endogenous Entry: Empirical Insights from Ebay Auctions," (2002), The Rand Journal of Economics
3. Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. Proceedings of the 11th Annual Conference on Computational Learning Theory (pp. 92-100).
4. Bryan, D., Lucking-Reily, D., Prasad, N., Reeves, D. Pennies from eBay: the Determinants of Price in Online Auctions., January 2000
5. Diettrich, T. & Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. Journal of Artificial Intelligence Research,2;263--286, 1995.

6. Oren Etzioni, Rattapoom Tuchinda, Craig A. Knoblock, Alexander Yates: To buy or not to buy: mining airfare data to minimize ticket purchase price. KDD 2003: 119-128
7. Ghani, R. Using Error-Correcting Codes for Text Classification. Proceedings of the 17th International Conference on Machine Learning (ICML 2000).
8. Ghani, R., Fano, A. Using Text Mining to Infer Semantic Attributes for Retail Data Mining. IEEE International Conference on Data Mining, 2002 (ICDM 2002)
9. JK MacKie-Mason, A Osepayshvili, DM Reeves, and MP Wellman. Price Prediction Strategies for Market-Based Scheduling. To appear, Fourteenth International Conference on Automated Planning and Scheduling, 2004.
10. Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning, 39, 103-134.
11. Robert E. Schapire, Peter Stone, David McAllester, Michael L. Littman, and János A. Csirik. Modeling Auction Price Uncertainty Using Boosting-based Conditional Density Estimation. In Proceedings of the Nineteenth International Conference on Machine Learning, 2002.
12. Wellman, M.P., Reeves, D.M., Lochner, K.M. and Vorobeychik, Y. (2004) "Price Prediction in a Trading Agent Competition", Journal of Artificial Intelligence Research, Volume 21, pages 19-36.